

Fast Multiplication of Matrices with Decay

Matt Challacombe and Nicolas Bock*

Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545[†]

A fast algorithm for the approximate multiplication of matrices with decay is introduced; the Sparse Approximate Matrix Multiply (SpAMM) reduces complexity in the product space, a different approach from current methods that economize within the matrix space through truncation or rank reduction. Matrix truncation (element dropping) is compared to SpAMM for quantum chemical matrices with approximate exponential and algebraic decay. For matched errors in the electronic total energy, SpAMM is found to require fewer to far fewer floating point operations relative to dropping. The challenges and opportunities afforded by this new approach are discussed, including the potential for high performance implementations.

INTRODUCTION

For large dense linear algebra problems, the computational advantage offered by fast matrix-matrix multiplication can be substantial, even with seemingly small gains in asymptotic complexity. Relative to conventional multiplication which is $\mathcal{O}(n^3)$, Strassen's algorithm achieves $\mathcal{O}(n^{2.8})$, while the Coppersmith and Winograd method is $\mathcal{O}(n^{2.38})$. For these dense methods, balancing the trade off between cost, complexity and error is an active area of research [1–3].

On the other hand, large sparse problems are typically handled with conventional sparse matrix techniques, with only small concessions between multiplication algorithms. Intermediate to these regimes, a wide class of problems exist that involve matrices with decay¹ where sparsity exists only asymptotically under an approximate linear algebra, historically involving matrix economization through element dropping or rank reduction. Often, problems with decay occur in the construction of matrix functions, notably the matrix inverse [6], the matrix exponential [7], and in the case of electronic structure theory, the Heaviside step function [4, 5, 8, 9]. The use of an approximate matrix algebra is also an active area of interest in the solution of large eigenproblems [10–13].

Many approaches to a sparse approximate linear algebra exist for matrices with decay [14–17], largely predicated upon the truncation of matrix elements, with the recent work of Benzi providing the most detailed analysis so far [4, 5]. In this contribution, we develop sparse matrix multiplication as a generalized N -body problem

[18], and introduce a fast algorithm based on hierarchical truncation in the three-dimensional space $i, j, k \in [1, n]$ of the product $C_{ij} = \sum_k A_{ik} B_{kj}$, where A and B decay exponentially or algebraically fast enough². Viewing the product from a length scale perspective¹, if matrix elements decay as $\mathcal{O}(1/r^\lambda)$, then the bulk of the product interactions will decay as $\mathcal{O}(1/r^{2\lambda})$. For small λ , the difference between truncation in the matrix space and the product space may be significant.

A SPARSE APPROXIMATE MATRIX MULTIPLY

The quadtree matrix representation,

$$A^k = \begin{pmatrix} A_{11}^{k+1} & A_{12}^{k+1} \\ A_{21}^{k+1} & A_{22}^{k+1} \end{pmatrix}, \quad k = 0, \dots, k_{\max}, \quad (1)$$

is the basis for recursive matrix-matrix multiplication, $C^k = A^k \cdot B^k$. For conventional recursive multiplication, the operator “ \cdot ” is just the row-column product, while in fast multiplication, it represents an economized sequence of operations with reduced complexity and a more complicated error accumulation. In Reference [19], Bini and Lotti carried out a detailed error analysis for recursive matrix multiplication schemes, and derived component-wise bounds of the form

$$|\tilde{c}_{ij} - c_{ij}| < a b \epsilon n \lg_2 n \quad (2)$$

where \tilde{c}_{ij} is a matrix element computed to within precision ϵ , c_{ij} is its exact counterpart, $a = \max_{ij} |a_{ij}|$ and $b = \max_{ij} |b_{ij}|$. While providing a sharp bound, the max norm does not immediately lend itself to the recursive separation of interaction magnitudes. Consider

¹ A matrix A is said to decay when its matrix elements decrease exponentially, as $|a_{i,j}| < c\lambda^{|i-j|}$, or algebraically as $|a_{i,j}| < \frac{c}{|i-j|^{\lambda+1}}$ with indicial separation $|i-j|$. In non-synthetic cases, the separation $|i-j|$ typically corresponds to an underlying physical distance $|\vec{r}_i - \vec{r}_j|$, *e.g.* of basis functions, finite elements, *etc.* See Figure 3 as well as the excellent work by Benzi and co-authors on this topic in References [4, 5].

² Algebraic decay sufficient to achieve a fast $\mathcal{O}(n \lg n)$ or better complexity is an open question similar to that of conditional convergence in the shape dependent summation of dipole and quadrupole components of the Lorentz field.

instead the framework provided by an arbitrary sub-multiplicative matrix norm $\|\cdot\|$:

$$\begin{aligned} \|C^k\| &\leq \|A^k\| \|B^k\| \\ &\leq \|A_{11}^{k+1}\| \|B_{11}^{k+1}\| + \|A_{12}^{k+1}\| \|B_{21}^{k+1}\| \\ &\quad + \|A_{11}^{k+1}\| \|B_{12}^{k+1}\| + \|A_{12}^{k+1}\| \|A_{22}^{k+1}\| \\ &\quad + \|A_{21}^{k+1}\| \|B_{11}^{k+1}\| + \|A_{22}^{k+1}\| \|B_{21}^{k+1}\| \\ &\quad + \|A_{21}^{k+1}\| \|B_{12}^{k+1}\| + \|A_{22}^{k+1}\| \|A_{22}^{k+1}\|. \end{aligned} \quad (3)$$

This structure suggests an algorithm we call the Sparse Approximate Matrix Multiply (SpAMM), which recursively tests each of the 8 contributions in Equation (3) for significance in accordance with a given numerical threshold τ :

$$\text{SpAMM}(A^k, B^k) = \begin{cases} 0 & \text{if } \|A^k\| \|B^k\| < \tau \\ A^k \cdot B^k & \text{elseif } k = k_{\max} \\ \left(\begin{array}{l} \text{SpAMM}(A_{11}^{k+1}, B_{11}^{k+1}) \\ + \text{SpAMM}(A_{12}^{k+1}, B_{21}^{k+1}) \\ \\ \text{SpAMM}(A_{21}^{k+1}, B_{11}^{k+1}) \\ + \text{SpAMM}(A_{22}^{k+1}, B_{21}^{k+1}) \end{array} \right) & \text{else} \end{cases} \quad (4)$$

The truncated product space accessed by SpAMM is shown in Figure 1 for matrices with exponential and algebraic decay.

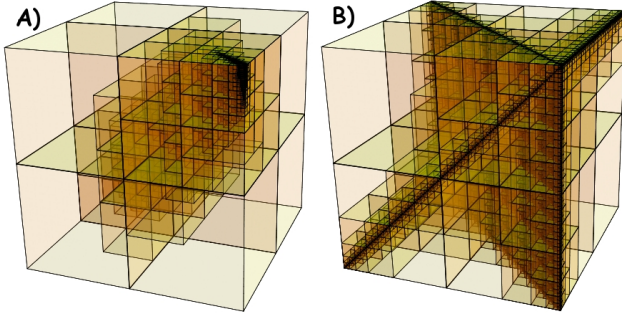


Figure 1: Hierarchical truncation of the product space (i, j, k) using $\tau = 10^{-8}$ for synthetic matrices of dimension $n = 512$, with $A_{ij} = \exp(-|i-j|)$ and $B_{ij} = \exp(-2|i-j|)$ in **(A)**, $A_{ij} = B_{ij} = \begin{cases} 1/|i-j|^3 & i \neq j \\ 0 & \text{else} \end{cases}$ in **(B)**. Each box above the finest ($k = k_{\max}$) scale represents truncation.

At each tier in SpAMM, the local truncation error is bounded by $\|\tilde{C}^k - C^k\| < \tau$, with an error accumulation structurally similar to rounding in conventional recursive multiplication, except that truncation is not guaranteed to occur at each tier in the recursion, and truncation does not retain even the approximate magnitude of the avoided sub-product. To see the difference between truncation and round-off, consider the generic, norm-wise

bound for recursive multiplication employed by Demmel and co-workers in Reference [2]:

$$\|\tilde{C} - C\| < \mu(n) \epsilon \|A\| \|B\| + \mathcal{O}(\epsilon^2), \quad (5)$$

with $\mu(n)$ a low order polynomial $\sim n^d$ and $d \geq 1$. Unlike rounding which is an error of commission, SpAMM creates errors of omission. If $\tau < \|A\| \|B\|$, then no work is performed and we obviously have $\|\tilde{C} - C\| < \tau$, which is different than $\|\tilde{C} - C\| < \epsilon \|A\| \|B\|$ corresponding to the case of comparable round-off and truncation parameters $\epsilon \sim \tau$. One could certainly bound SpAMM rigorously by decreasing τ with increasing depth, $\tau^{k+1} = \tau^k/8$, but that would be overly pessimistic, not taking into account signed error accumulation or the localization and attenuation of errors due to decay.

SpAMM is similar to the \mathcal{H} -matrix algebra of Hackbusch and co-workers, where off-diagonal sub-matrices are treated as reduced rank factorizations (truncated SVD), typically structured and grouped to reflect properties of the underlying operators [20]. For problems with rapid decay, truncated SVD may behave in a similar way to simple dropping schemes. SpAMM is different than the \mathcal{H} -matrix algebra as it achieves separation uniquely in the product space and does not rely on a reduced complexity representation of matrices. For very slow decay, the \mathcal{H} -matrix algebra may certainly offer a path for intractable problems for which SpAMM is ineffective.

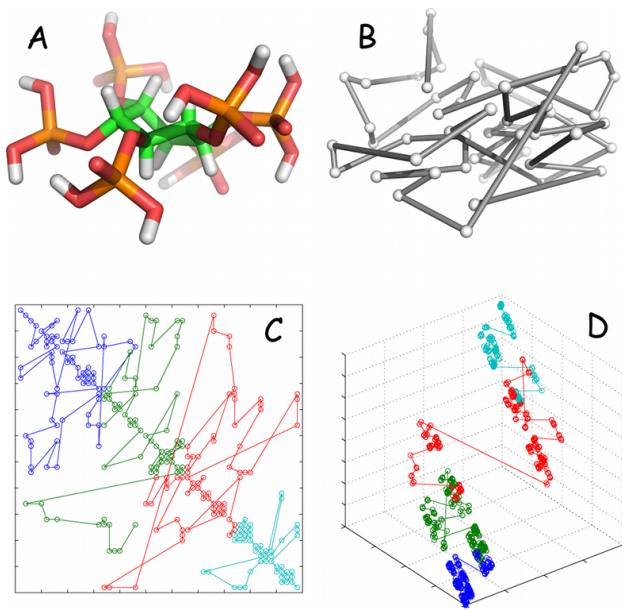


Figure 2: Space filling curves map atoms in Cartesian space (A) onto an index that is locality preserving (B), leading to clustering of matrix elements with respect to indicial separation (C) when interactions are short ranged. The octree generated by SpAMM in the three-dimensional product space, shown in Figure 1, is equivalent to a second space filling curve (spatial hashing) [21–23] that orders both matrices (C) and the product space (D), features that can be exploited to achieve domain decomposition and load balance (colors in C and D).

An understanding of the error accumulation in SpAMM must account for the decay properties of A and B , which in non-synthetic cases are intimately related to the effects of ordering and structure of the underlying physical, chemical or engineering application. Also, ordering will determine the relative efficiencies of both matrix truncation and SpAMM, particularly under blocking. The ordering used here is based on the Space Filling Curve (SFC) method described in [9] (Hilbert atom ordering), and shown in Figure 2, involving also a second tier of ordering in the product space. The first ordering maps atoms that are close in Cartesian-space to entries close in the index space of the matrix. The second ordering is a natural consequence of the SpAMM multiply, which recursively maps out a multi-level octree (see also Figure 1) with cuboid coordinates that are equivalent to a spatial-hash; sorting the hash produces a three-dimensional SFC in the product space [21–23]. This two-tiered structure is novel, providing an encompassing scheme that parleys the locality of physical interactions into the data locality of matrices (element clustering), and into the irregular product space. This approach is applicable to other problems with underlying decay properties, in which the finite-elements, radial basis functions, *etc.* replace atoms, or graph theoretical methods (nested dissection, Cuthill-McKee, *etc.*) replace

the first-tier SFC altogether. In either case, the correspondence between the recursive SpAMM product space and a three-dimensional SFC provides a tool to exploit both spatial and temporal locality in the distribution of work and data.

RESULTS AND DISCUSSION

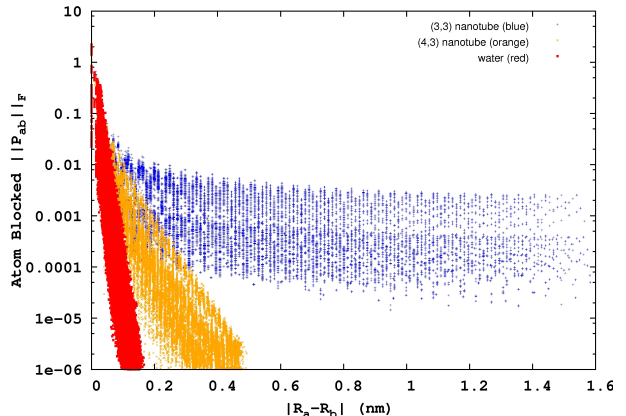


Figure 3: Decay of normed density matrix atom-blocks $\|P_{ab}\|_F$ with Cartesian separation $|\vec{R}_a - \vec{R}_b|$ for the largest molecules in each sequence: a 450 atom water cluster (red), a 752 atom (4,3) nanotube (orange) and a 780 atom (3,3) nanotube (blue).

In this initial contribution, we limit ourselves to exploring the numerical and computational behavior of SpAMM applied to problems in electronic structure theory, and compare its relative merits to the dropping of matrix elements. In $\mathcal{O}(n)$ electronic structure calculations [15–17], a primary source of error due to sparse matrix-multiplication develops from early steps in the iterative construction of the Heaviside matrix function $P = \theta[F - \mu I]$ (density matrix purification), which is a projector of the effective Hamiltonian (Fockian) F [24]. Starting from the basis spanning F , purification drives eigenvalues to 0s or 1s, whereupon error accumulation due to an approximate linear algebra is quenched [24]. Under a given approximate linear algebra, the electronic energy $E_{el} = \text{Tr}(P.F)$ is a global measure of accumulated error. In the following Section, we carry out purification on three molecular sequences of increasing size, water clusters, (4,3) nanotubes and (3,3) nanotubes. In each case, a Fockian F was obtained from a fully converged Self-Consistent-Field (SCF) cycle and used as basis for the purification; our numerical experiments probe only errors within one density matrix solve (40-60 multiplies), and do not address error propagation throughout the SCF cycle. As the rate of decay slows towards SCF convergence, these calculations represent the instance of

minimum decay, shown for each of the largest molecular species in Figure 3. Within each sequence, the same number of iterative steps (40-60 matrix multiplications) are taken using the TC2 purification algorithm [25] and either: (I) matrix element dropping and exact multiplication or (II) SpAMM.

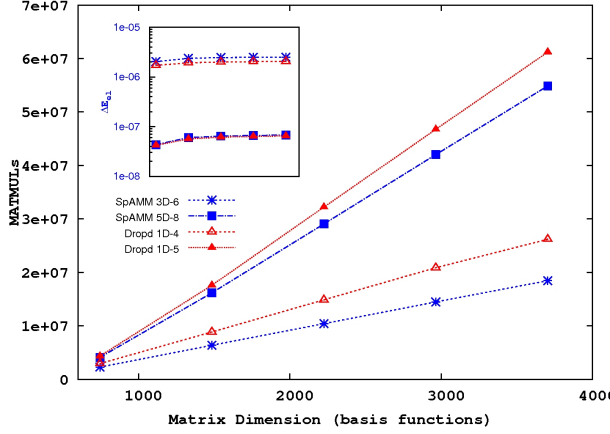


Figure 4: Average number of 4x4 matrix multiplies (MATMULs) for a sequence of (4,3) nanotubes at the RHF/STO-2G level of theory with dropping (red) and SpAMM (blue).

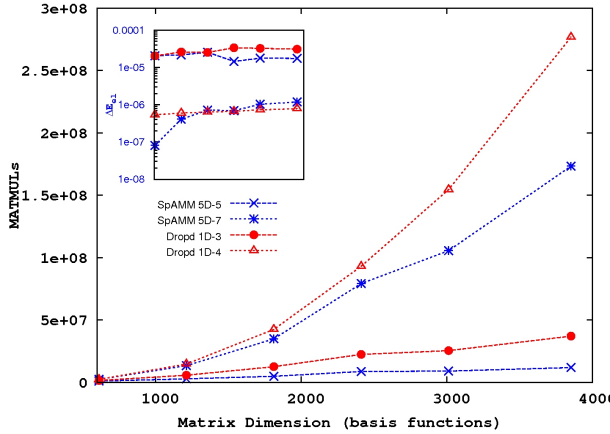


Figure 5: Average number of 4x4 matrix multiplies (MATMULs) for a sequence of (3,3) nanotubes at the LDA/STO-2G/ level of theory with dropping (red) and SpAMM (blue).

In this study, we chose k_{\max} to yield 4x4 blocks at the finest level of resolution, corresponding to the most aggressive use of single precision SSE on the x86 architecture. In all cases the matrix norm employed is the Frobenius norm $\|\cdot\| \equiv \|\cdot\|_F$. Thresholds, τ , have been adjusted to roughly match relative errors in the electronic energy, $\Delta E_{\text{el}} = |\tilde{E}_{\text{el}} - E_{\text{el}}|/|E_{\text{el}}|$ between the two schemes, (I) and (II), and the average number of 4x4 MATMULs per purification step are reported. Multiplications are only a proxy for CPU time, as neither case accounts for symbolic overheads associated with the multiply (CSR,

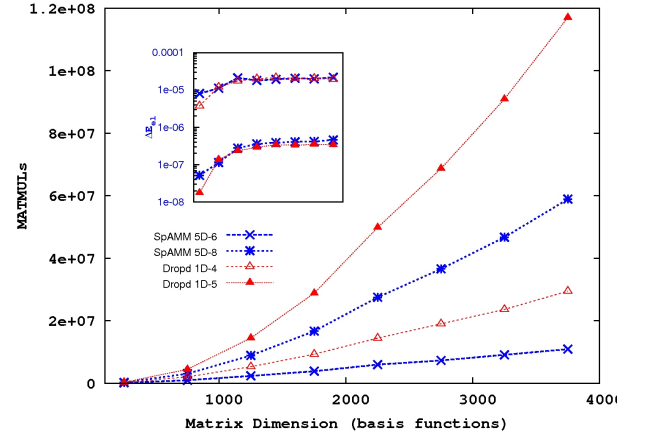


Figure 6: Average number of 4x4 matrix multiplies (MATMULs) for a sequence of water clusters at the RHF/6-31G** level of theory with dropping (red) and SpAMM (blue).

recursive-tree *etc.*). In the case of dropping, SpAMM was also used in the matrix multiply but with zero threshold. After each multiplication in the dropping scheme, a filter was applied to the resultant, dropping blocks at the 4x4 level of resolution using the criteria $\|P^k\|_F < \tau$. Results for the three molecular sequences are shown in Figures 6, 4 and 5.

For comparable values of ΔE_{el} , SpAMM was found to employ from slightly fewer multiplies for the (4,3) nanotube, to dramatically less in the case of the water clusters. This result is somewhat surprising, since the spatial decay is slower in the case of the (4,3) nanotube than for the water clusters, as shown in Figure 3. While it seems reasonable to attribute this unexpected result to the effects of dimensionality, further study is required to be sure. Its also worth noting that the advantage of SpAMM relative to dropping is brought down with decreasing τ ; for $\tau = 0$ they both revert to the same $\mathcal{O}(n^3)$ complexity. For both systems with exponential decay, error appears tightly controlled albeit within an as yet unknown bound. Note however, that unlike matrix truncation which leads to a product error that is $\mathcal{O}(\tau^2)$, SpAMM leads to a truncation error that is $\mathcal{O}(\tau)$.

Comparing the quasi one-dimensional metallic (3,3) nanotube with slow algebraic decay to the insulating (4,3) nanotube with exponential decay, SpAMM gains substantially over dropping in the case of slower decay. However, the ability of SpAMM to achieve a linear scaling complexity in the case of the metallic system remains in question, as in the case of the tightest threshold, the SpAMM errors do not appear to be well controlled, and the cost does not appear linear, at least not in this size regime. On the other hand, the SpAMM result does enjoy a significant reduction in cost relative to dropping, and the error increase is modest.

CONCLUSION

The Sparse Approximate Matrix Multiply (SpAMM) is a fast method for matrices with decay, which is different from element dropping or the \mathcal{H} -algebra in that it uniquely involves truncation of the product space rather than the matrix space. For matrices with exponential or fast algebraic decay, SpAMM can achieve stable error control comparable to element dropping, but with a greatly reduced number of floating point operations. The results presented here are preliminary, and have not yet explored the interesting problems of slow decay in the asymptotic limit, ordering, error bounds, the cost of recursion, high performance implementations or broader gauges of accuracy and efficiency, such as the use of SpAMM in the context of the Self-Consistent-Field cycle. Of particular concern is the relationship between complexity, matrix decay and error control. Based on our numerical tests, we postulate that the algorithm is at worst $\mathcal{O}(n \lg n)$ for matrices with sufficiently fast decay.

In addition to similarities with the \mathcal{H} -algebra, SpAMM falls under the rubric of the generalized N -body problem [18]. From this perspective, it is worth noting also the connection between matrix-matrix multiplication as N -body problem and matrix-matrix multiplication as spatial join [26], as well as between N -body problems and data base theory in general [23].

Next, we draw attention to the second tier of Space Filling Curve (SFC) shown in Figure 2, which provides a mechanism for domain decomposition and load balance that is proven for parallel irregular problems [21, 22, 27, 28]. Also, the improvement gained in Reference [29] on going from a one-dimensional to a two-dimensional matrix partitioning scheme for the parallel SpMM suggests that partitioning the three-dimensional product space instead may provide an even higher degree of flexibility and granularity.

The authors acknowledge support through Los Alamos LDRD award ER20110230 (computational co-design) as well as funds from the U.S. Department of Energy. Los Alamos National Laboratory is operated by the Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under Contract No. DE-AC52-06NA25396. Special acknowledgments go to the International Ten Bar Café for tasty libations in a scientific and collegial atmosphere, and to Michele Benzi for valuable input.

* Electronic address: mchalla,nbock@lanl.gov

† URL: freeon.org

[1] J. W. Demmel and N. J. Higham, ACM Trans. Math. Softw. **18**, 274 (1992).

- [2] J. Demmel, I. Dumitriu, O. Holtz, and R. Kleinberg, Numer. Math. **106**, 199 (2007).
- [3] R. Yuster and U. Zwick, ACM Transactions on Algorithms **1**, 2 (2005).
- [4] M. Benzi and N. Razouk, Electronic Transactions on Numerical Analysis **28**, 16 (2007).
- [5] M. Benzi, P. Boito, and N. Razouk, Manuscript in preparation (2010).
- [6] M. Benzi and M. Tuma, SIAM Journal on Scientific Computing **21**, 1851 (2000).
- [7] A. Iserles, *How large is the exponential of a banded matrix?* (1999).
- [8] M. Challacombe, J. Chem. Phys. **110**, 2332 (1999).
- [9] M. Challacombe, Computer Physics Communications **128**, 93 (2000).
- [10] V. Simoncini and L. Elden, Bit Numerical Mathematics **42**, 159 (2002).
- [11] V. Simoncini and D. B. Szyld, SIAM Journal on Scientific Computing **25**, 454 (2003).
- [12] V. Simoncini and D. B. Szyld, SIAM Review **47**, 247 (2005).
- [13] M. Challacombe, arXiv **quant-ph**, 1001.2586 (2010).
- [14] G. Galli, Current Opinion in Solid State & Materials Science **1**, 864 (1996).
- [15] S. Goedecker, Reviews of Modern Physics **71**, 1085 (1999).
- [16] S. Goedecker and G. Scuseria, Computing in Science Engineering **5**, 14 (2003).
- [17] Z. Y. Li, W. He, and J. L. Yang, Progress in Chemistry **17**, 192 (2005).
- [18] A. G. Gray and A. W. Moore, in *Advances in Neural Information Processing Systems* (MIT Press, 2001), vol. 4, pp. 521–527.
- [19] D. Bini and G. Lotti, Numerische Mathematik **36**, 63 (1980).
- [20] L. Grasedyck and W. Hackbusch, Computing **70**, 2003 (2003).
- [21] M. S. Warren and J. K. Salmon, in *Supercomputing '92* (IEEE Comp. Soc., Los Alamitos, 1992), pp. 570–576, (1992 Gordon Bell Prize winner).
- [22] M. S. Warren and J. K. Salmon, *A parallel, portable and versatile treecode* (SIAM, Philadelphia, 1995), chap. 1.
- [23] H. Samet, *Foundations of Multidimensional and Metric Data Structures* (Morgan Kaufmann, 2006).
- [24] A. M. N. Niklasson, C. J. Tymczak, and M. Challacombe, J. Comp. Phys. **118**, 8611 (2003).
- [25] A. M. N. Niklasson, Physical Review B **66**, 5 (2002).
- [26] R. Amossen and R. Pagh, in *Proceedings of the 12th International Conference on Database Theory* (ACM, 2009), pp. 121–126.
- [27] S. Aluru and F. E. Sevilgen, in *Proceedings of the 4th IEEE Conference on High Performance Computing* (1997), pp. 230–235.
- [28] K. D. Devine, E. G. Boman, R. T. Heaphy, B. A. Hendrickson, J. D. Teresco, J. Faik, J. E. Flaherty, and L. G. Gervasio, Applied Numerical Mathematics **52**, 133 (2005), ADAPT '03: Conference on Adaptive Methods for Partial Differential Equations and Large-Scale Computation.
- [29] A. Buluc and J. R. Gilbert, in *ICPP '08: Proceedings of the 2008 37th International Conference on Parallel Processing* (IEEE Computer Society, Washington, DC, USA, 2008), pp. 503–510.